Support Vector Machines
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
Binary Classification Problem

- **Training data**: sample drawn i.i.d. from set \( X \subseteq \mathbb{R}^N \) according to some distribution \( D \),

\[
S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in X \times \{-1, +1\}.
\]

- **Problem**: find hypothesis \( h : X \mapsto \{-1, +1\} \) in \( H \) (classifier) with small generalization error \( R_D(h) \).

- **Linear classification**:
  - Hypotheses based on hyperplanes.
  - Linear separation in high-dimensional space.
Classifiers: \( H = \{ x \mapsto \text{sgn} (w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R} \} \).
Optimal Hyperplane: Max. Margin

(Vapnik and Chervonenkis, 1965)

**Canonical hyperplane:** \( w \) and \( b \) chosen such that for closest points \( |w \cdot x + b| = 1 \).

**Margin:** \( \rho = \min_{x \in S} \frac{|w \cdot x + b|}{\|w\|} = \frac{1}{\|w\|} \).
Optimization Problem

Constrained optimization:

\[
\min_{w,b} \frac{1}{2} \|w\|^2
\]

subject to \( y_i(w \cdot x_i + b) \geq 1, i \in [1, m] \).

Properties:

• Convex optimization (strictly convex).
• Unique solution for linearly separable sample.
Optimal Hyperplane Equations

- **Lagrangian**: for all $w, b, \alpha_i \geq 0$,

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1].$$

- **KKT conditions**:

$$\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i.$$  
$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0.$$  
$$\forall i \in [1, m], \alpha_i [y_i(w \cdot x_i + b) - 1] = 0.$$
Support Vectors

- Complementarity conditions:

\[ \alpha_i [y_i (w \cdot x_i + b) - 1] = 0 \implies \alpha_i = 0 \lor y_i (w \cdot x_i + b) = 1. \]

- Support vectors: vectors \( x_i \) such that

\[ \alpha_i \neq 0 \land y_i (w \cdot x_i + b) = 1. \]

- Note: support vectors are not unique.
Moving to The Dual

- Plugging in the expression of $w$ in $L$ gives:

$$L = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i. $$

- Thus,

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).$$
Dual Optimization Problem

- **Constrained optimization:**

\[
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j (x_i \cdot x_j)
\]

subject to: \(\alpha_i \geq 0 \land \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]\).

- **Solution:**

\[
h(x) = \text{sgn}\left( \sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b \right),
\]

with \(b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)\) for any SV \(x_i\).
Leave-One-Out Analysis

**Theorem:** let $h_S$ be the optimal hyperplane for a sample $S$ and let $N_{SV}(S)$ be the number of support vectors defining $h_S$. Then,

$$E_{S \sim D^m}[R(h_S)] \leq E_{S \sim D^{m+1}} \left[ \frac{N_{SV}(S)}{m + 1} \right].$$

**Proof:** Let $S \sim D^{m+1}$ be a sample linearly separable and let $x \in S$. If $h_{S-\{x\}}$ misclassifies $x$, then $x$ must be a SV for $h_S$. Thus,

$$\hat{R}_{1oo}(\text{opt.-hyp.}) \leq \frac{N_{SV}(S)}{m + 1}.$$
Notes

- Bound on expectation of error only, not the probability of error.
- Argument based on sparsity (number of support vectors). We will see later other arguments in support of the optimal hyperplanes based on the concept of margin.
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
Support Vector Machines

(Cortes and Vapnik, 1995)

- **Problem**: data often not linearly separable in practice. For any hyperplane, there exists $x_i$ such that

\[ y_i [w \cdot x_i + b] \geq 1. \]

- **Idea**: relax constraints using slack variables $\xi_i \geq 0$

\[ y_i [w \cdot x_i + b] \geq 1 - \xi_i. \]
Soft-Margin Hyperplanes

- **Support vectors**: points along the margin or outliers.
- **Soft margin**: $\rho = \frac{1}{\|w\|}$. 

Mathematical expressions:

\[ w \cdot x + b = 0 \]
\[ w \cdot x + b = +1 \]
\[ w \cdot x + b = -1 \]

\[ \xi_i \]
\[ \xi_j \]

\[ \rho = \frac{1}{\|w\|} \]
Optimization Problem

(Cortes and Vapnik, 1995)

Constrained optimization:

\[
\min_{w, b, \xi} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i(w \cdot x_i + b) \geq 1 - \xi_i \land \xi_i \geq 0, i \in [1, m] \).

Properties:

- \( C \geq 0 \) trade-off parameter.
- Convex optimization (strictly convex).
- Unique solution.
Parameter $C$ : trade-off between maximizing margin and minimizing training error. How do we determine $C$?

The general problem of determining a hyperplane minimizing the error on the training set is NP-complete (as a function of dimension).

Other convex functions of the slack variables could be used: this choice and a similar one with squared slack variables lead to a convenient formulation and solution.
Hinge Loss

‘Quadratic’ hinge loss $\xi^2$

Hinge loss $\xi^1$

0/1 loss function
SVMs Equations

- **Lagrangian**: for all \( w, b, \alpha_i \geq 0, \beta_i \geq 0, \)

\[
L(w, b, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i \xi_i.
\]

- **KKT conditions**:

\[
\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i.
\]

\[
\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0.
\]

\[
\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \iff \alpha_i + \beta_i = C.
\]

\[\forall i \in [1, m], \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0\]

\[\beta_i \xi_i = 0.\]
Support Vectors

Complementarity conditions:

\[ \alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0 \implies \alpha_i = 0 \lor y_i(w \cdot x_i + b) = 1 - \xi_i. \]

Support vectors: vectors \( x_i \) such that

\[ \alpha_i \neq 0 \land y_i(w \cdot x_i + b) = 1 - \xi_i. \]

Note: support vectors are not unique.
Moving to The Dual

- Plugging in the expression of $w$ in $L$ gives:

$$L = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i.$$

- Thus,

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).$$

- The condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$. 

Mehryar Mohri - Introduction to Machine Learning
Dual Optimization Problem

Constrained optimization:

\[
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)
\]

subject to: \(0 \leq \alpha_i \leq C \land \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].\)

Solution:

\[
h(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right),
\]

with \(b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)\) for any \(x_i\) with \(0 < \alpha_i < C\).
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
Margin Loss

Definition: for any $\rho > 0$, the $\rho$-margin loss is the function $L_\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ defined by for all $y, y' \in \mathbb{R}$ by

$$L_\rho(y, y') = \Phi_\rho(yy')$$

with

$$\Phi_\rho(x) = \begin{cases} 
0 & \text{if } \rho \leq x \\
1 - x/\rho & \text{if } 0 \leq x \leq \rho \\
1 & \text{if } x \leq 0.
\end{cases}$$

For a sample $S = (x_1, \ldots, x_m)$ and a hypothesis $h$, the empirical loss is

$$\hat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) < \rho}.$$
**Margin Bound - Linear Classifiers**

**Corollary:** Let $\rho > 0$ and $H = \{ x \mapsto w \cdot x : \|w\| \leq \Lambda \}$. Assume that $X \subseteq \{ x : \| x \| \leq R \}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}_\rho(h) + 2 \sqrt{\frac{R^2 \Lambda^2 / \rho^2}{m}} + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$
High-Dimensional Feature Space

- **Observations:**
  - generalization bound does not depend on the dimension but on the margin.
  - this suggests seeking a large-margin separating hyperplane in a higher-dimensional feature space.

- **Computational problems:**
  - taking dot products in a high-dimensional feature space can be very costly.
  - solution based on *kernels* (next lecture).
References


